# Comment

**Sanford Weisberg**

Professors Chatterjee and Hadi examine the algebraic representations and relationships between many of the statistics that have been proposed to monitor "influence." For a thorough understanding of influential data, perhaps it may be more useful to study the underlying principles used in defining the measures, since the algebra can obscure the basic ideas. For example, in their Section 5.2, Chatterjee and Hadi describe "the" likelihood displacement by formula (23). If $L(\beta, \sigma^2)$ is the log likelihood surface for $\beta$ *and* $\sigma^2$, then the quantity given by (23) is $2[L(\hat{\beta}, \hat{\sigma}^2) - L(\hat{\beta}_{(i)}, \hat{\sigma}^2_{(i)})]$, or twice the difference in height of this surface at the full data *mle* and at the *mle* with case $i$ deleted. It incorporates changes in the estimates of both $\beta$ and $\sigma^2$ in a natural way, a fact that is surely not evident from (23). Thus it is different from most of the other influence measures discussed in the paper that are concerned either with $\beta$ or with $\sigma^2$ alone, or use some ad hoc method to combine the effects of changes in both of them. If one considers the log likelihood for $\beta$ alone by maximizing over $\sigma^2$, a different measure will result that is a monotonic function of Cook's distance. Not everyone will agree that changes in the height of a likelihood surface is an important criterion for measuring influence, but this approach does give specific meaning to the idea of an influential case and can therefore provide a basis for further work.

Influence analysis must depend at least in part upon the context in which it is applied. Consider again the example used in the paper. The data were collected by Jim Moore, an Agricultural Engineer then at the University of Minnesota, to study the biological activity in dairy wastes as it decomposes. He suspended a sample in a ceramic crock, and measured biological activity for about 220 days. Without looking at the data, one should expect that 1) biological activity will be very high at the beginning of the time period (case 1); 2) eventually, biological activity will cease, so after day $t$, where $t$ may be less than 220 (case 20), the response, not its logarithm, will be zero or at least essentially constant except for measurement error; and 3) the response, and probably the predictors as well, should be decreasing but probably nonlinear functions of time or case number. Taking log(oxygen demand) as response and fitting linear models is a reasonable first step. If we do so, we should expect that results on day 1 may be much more variable than will observations later in the experiment, while days at the end of the experiment will have smaller residual variance and, after some day $t$, the log(response) is likely to be essentially constant. What can diagnostics tell us about these data, given this preamble?

As a simple but powerful approach, suppose we follow the advice given in Weisberg (1985, Chapter 5): examine $h_i$, $C_i$, and either of $t_i$ or $t_i^*$. We immediately find case 1 to have a comparatively large residual, and moderate values for the other two diagnostics. This agrees with the expectation of large error variation at the beginning of the experiment. Similarly, day 220, case 20, is a clear problem, with large $C_{20}$ and large $h_{20}$. The complete data, including day numbers, are given by Weisberg (1985, page 222). By consulting this source, we see that oxygen uptake on day 220 was measured as 0.9, while the preceding six measurements were 0.4, 0.6, 0.7, 0.6, 0.4, and 0.3 for days 100–171, respectively. While we cannot be sure when activity became essentially constant, since it should be nonincreasing with time, it seems to have stabilized before the end of the experiment. The comparatively large reading of 0.9 is probably due to measurement error.

The other "interesting" case in these data is number 17—day number 129, with very large values for all the diagnostic statistics. This unexpected finding is new information that might have been missed, and we may wish to find the cause. Table 3 in the paper (*the raw data*) contains the answer: there is a typographical error in case 17, the number 57.7 for total volatile solids, should be 75.7 (the error is in Moore's thesis and I have preserved it in my book as an interesting exercise for students). The error should be fairly obvious since total volatile solids will decrease with time, and the value given for case 17 clearly does not fit the pattern. If this is corrected, $C_{17}$ decreases from 1.78 to only 0.0003. These seem to be the major features of this data set, and they can be found by using three simple building blocks, $h_i$, $t_i$, and $C_i$, and perhaps a little common sense.

With nearly 10 years experience using influence methods in the linear regression problem, we should be able to describe methods using a general framework such as the likelihood displacement that can include other statistical problems. Unfortunately, by stressing algebra over philosophy, geometry, theory, and

*Sanford Weisberg is Professor of Statistics, Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55108.*

context, this paper does not provide the reader with the necessary ideas. Although many papers have been published in this area since 1982, only two papers published since then are cited. Discussion of many of the fundamental issues, including a comprehensive review of the literature to 1982, can be found in Cook and Weisberg (1982, especially Section 5.2); see also Weisberg (1983) and Cook (1986).

## ADDITIONAL REFERENCES

COOK, R. D. (1986). Assessment of local influence (with discussion). To appear in *J. Roy. Statist. Soc. Ser. B* **48**.

WEISBERG, S. (1983). Some principles for regression diagnostics and influence analysis, discussion of a paper by R. R. Hocking. *Technometrics* **25** 240–244.

WEISBERG, S. (1985). *Applied Linear Regression*, 2nd ed. Wiley, New York.

# Rejoinder

**Samprit Chatterjee and Ali S. Hadi**

Many points have been raised, but alas, space does not permit us to respond to each one of them individually. For expedience, the comments which we feel have arisen due to a misreading of what we wrote will not be discussed, letting the readers make up their own minds. Our paper will have served its purpose if it stimulates discussion and leads to further development in methodology. We are grateful to Professor DeGroot for getting together such a distinguished group to act as discussants for our paper.

Several of the authors (Brant, Hoaglin and Kempthorne, and Welsch), have pointed out very correctly that little was said in our paper about detecting groups or clusters of influential points. Not much is known; and we came to know about the work of Brant and Kempthorne only recently. We are not convinced, however, as to how real the problem is. Most of the influential points may be detected by a one point at a time deletion scheme. We see our skepticism on this point is also shared by Welsch.

Weisberg has noted that we have not provided an overriding general principle for deriving various influence measures. Space considerations prevented such an effort. Basically we tend to favor the influence function approach introduced by Hampel. In our forthcoming book, *Sensitivity Analysis in Linear Regression*, we outline such an approach. We show that almost all proposed influence measures can be derived from various approximations of the influence function. The likelihood approach, as pointed out by Cook and Weisberg, is another unifying principle. We are not convinced, however, about its robustness. We prefer measures which are based on metric distances rather than those based on probability densities, and therefore we have stayed away from influence measures based on information theory.

Several authors have raised questions about the callibration points which we have provided in Table 2. There is nothing sacred about them. They are meant to be yardsticks, equivalent to "± 2 standard error rules." Our attitude to them is identical with those articulated by Velleman and Hoaglin and Kempthorne, although it might not have been stated as explicitly. Points which stand out from the group on their diagnostic measures should certainly be flagged and examined. It is the standing apart which should trigger off the alarm rather than the exceeding of a critical value. Stem and leaf plots are very effective graphical devices for this purpose. We would like to endorse the diagnostic strategy advocated by Hoaglin and Kempthorne. In fact it is this approach which has led us to flag points 1 and 17 on the basis of $\mathrm{CVR}_i$ rather than all the points which mechanically lie outside the critical interval. Points in Table 5 are starred only when they stand out (outliers on the diagnostic measure) rather than merely exceed their calibration values. We thank the discussants for highlighting this point.

Several of the discussants brought up the important question of observations influencing variable selection in model determination. Most influence measures do not distinguish whether an observation is influential on all dimensions or only on one or few dimensions. An observation, for example, might appear to be the most influential one according to a given measure, but when a particular variable is omitted the influence disappears. Retaining a variable may hinge on one or a few observations. In our present paper, we did not discuss this complex question, but have a paper languishing somewhere in the refereeing process, which addresses this question. The role of observations in variable selection (irrespective of the criteria used) is an area which needs clarification.

Atkinson in his related comments makes a point not made by the other discussants. If we have read his comments correctly, it appears that he opts for a robust estimation procedure. This is certainly a valid approach. A model fitting approach in which no point